The Effects of Feedback on Item-by-Item and Global Judgments of Learning

Rafal Kozlowski

A report submitted as a partial requirement for the degree of Bachelor of Psychological Science with Honours at the University of Tasmania, 2017

Statement of Sources

I declare that this report is my own original work and that contributions of others have been duly acknowledged.

Rafal Kozlowski

Date:

Acknowledgments

Above all, I am grateful to my supervisor, Dr Jim Sauer, for not only his deluxe academic support but also for his ability to lift my spirits during the difficult moments.

My gratitude extends to Dr Matt Palmer and Laura Brumby for finding time to assist me in recruiting participants, running experiments and retrieving data.

I'd especially like to thank my best friend who supported me throughout the whole Honours process – my wife Kasia: what doesn't kill us makes us stronger!

A special thank you also to my children, Tobiasz and Klara, for their understanding, patience and unwavering belief in me.

I wish to also thank my parents, for their encouragement of my studies and their love and support throughout my life, and my parents-in-law, for their help, particularly with the children, in this busy year.

Table of Contents

List of Tables and Figures

The Effects of Feedback on Item-by-Item and Global Judgments of Learning

Rafal Kozlowski

Word count: 10000

Abstract

This study examined the effect of accurate feedback and biased feedback (inflated and deflated), in the form of marks, on the accuracy of item-by-item and global judgments of learning (JOL). 80 participants (49 females, M_{age} = 29.61, SD = 12.76) were randomly allocated into one of four conditions: no feedback, accurate feedback, inflated feedback and deflated feedback. Using a computer program, participants studied 50 Swahili-English word-pairs, and judged the likelihood of remembering each item (item-by-item JOL) and the overall percentage of likely remembered word-pairs after each learning session (global JOL). Immediately after each learning session, a testing session was held. The accurate feedback group received accurate marks after the first and second testing session. The inflated feedback group received marks increased by 32 percent; the deflated feedback group – marks decreased by 32 percent. There were three study-test phases. Feedback did not affect item-by-item JOL accuracy. Deflated feedback significantly reduced global JOL accuracy with a large effect size (d = 1.16), suggesting that participants are sensitive to deflated feedback, which can potentially affect their theory-based cues about their ability to learn.

This research project investigated the effects of feedback on metacognition – the ability to judge one's own knowledge. This self-knowledge can be understood as higher order thinking, where the subject of this mental process is itself performing the thinking. Metacognition is widely considered a critical skill in educational settings (Dunlosky & Metcalfe, 2008; Metcalfe, 2009, 2017; Nelson, Dunlosky, Graf, & Narens, 1994) as metacognitive monitoring is thought to drive metacognitive control (e.g., the allocation of study time to items requiring further learning). Thus, effective monitoring is critical for effective behavioural regulation and, ultimately, learning (Dunlosky & Metcalfe, 2008). However, metacognitive knowledge is not uniform and does not yield homogenous effects on learning performance. Rather, its benefits depend on monitoring accuracy (Rawson, O'Neil, & Dunlosky, 2011) which in turn depends, for example, on the characteristics of the subject being studied, the conditions of studying, and the mnemonic processes involved (Koriat, 1997). We investigated how metacognitive judgments are affected by assessment feedback, comparing the effects of accurate feedback, and positively or negatively biased feedback, on the accuracy of predictions in subsequent studying cycles, against a no-feedback control condition.

Development of Metacognitive Research

The first notions of a higher order cognition date back to Aristotle's treatise *On the Soul* (trans. 1931) which stated that the mind itself can be an object of thinking in the same way as external objects. However, for radical behaviourists, prevalent in psychology in the middle of the last century, cognition was conceptualised as another kind of behaviour (Skinner, 1945) or as stimuli affecting behaviour (Razran, 1955). Therefore, cognition and metacognition became relevant in evidence-based psychology only when the behavioural approach started losing its

monopoly on rigorous quantitative research to the school of cognitive psychology (Staddon, 2014).

The term metacognition came into use when Flavell (1971) presented the idea that children, as they develop, increase their ability to monitor their knowledge, which he called "metamemory". Ultimately, Flavell (1976) conceptualised metamemory as a subtype of metacognition and defined metacognition as knowledge about one's own cognitions, commonly referred to as thinking about thinking, of which metamemory is a specific example relating to knowledge of one's own memory.

Judgments of Learning

A key measure of metamemory in comparatively recent studies has been judgments of learning (JOLs). People make JOLs during or after a learning session by predicting the likelihood of future recall (Dunlosky & Metcalfe, 2008; Nelson & Narens, 1990). In a typical JOL study, participants study a list of word-pairs consisting of a cue and a target word, for example, FORK (cue) – SPOON (target). They then rate their probability of being able to recall the target when presented only with the cue. This rating is their JOL. The ratings can then be compared with actual recall performance in a subsequent test to assess the accuracy (or predictive validity) of JOLs.

Researchers have conflicting views regarding the theoretical basis for JOLs. According to the early direct-access hypothesis, metacognitive monitoring judgments are made by directly accessing and indexing the strength of a target item in memory, e.g., through an attempt at retrieval (King, Zechmeister, & Shaughnessy, 1980; Koriat, 2000). Hence, the direct-access hypothesis is falsified when the outcome of memory monitoring diverges from the memory performance. This divergence has

been found in most JOL studies to date, which revealed that participants are not perfectly accurate when making JOLs (e.g. Koriat, Sheffer, & Ma'ayan, 2002). These studies revealed that although people can effectively monitor their learning, metacognitive judgments are vulnerable to systematic distortions. Therefore, it is possible to dissociate effects on metacognition from effects on memory performance.

An explanation for the disparity between predictions and memory performance was proposed by Koriat's (1997) cue-utilisation hypothesis. Koriat argued that when people produce JOLs they utilise simultaneously three types of cues: intrinsic, extrinsic, and mnemonic. According to this typology, intrinsic cues relate to the characteristics of a studied item (e.g., perceived difficulty). For example, closely associated word-pairs of high frequency, such as cat – milk, would cue low difficulty of learning whereas a distant association of low-frequency words, such as damsel – mitochondria, would indicate higher learning difficulty. Koriat proposed that the second class of cues – extrinsic – involves both the conditions of learning (e.g., the number of study trials) and the types of mental operations performed, (e.g., the degree of elaboration). According to the cue-utilisation approach, the analysis of both intrinsic and extrinsic cues can affect JOLs directly via a theory-based pathway. This pathway of judgment, according to Kelley and Jacoby (1996) relies on a pre-existing collection of rules which are used in the judgement process. The analysis of the theory-based pathway is focused on specified factors, for example, the difficulty of the subject being studied (intrinsic cue), and the amount of time spent studying (extrinsic cue).

Koriat (1997) proposed that people also use heuristics when making JOLs and that they use their past experience as a basis for these mental shortcuts. Koriat labelled this third class of cues mnemonic – intuitive judgments on the accessibility

of an item from memory. Koriat argued that mnemonic cues are not a mere reflection of memory tracing, but are also affected by intrinsic and extrinsic cues. Therefore, according to the cue-utilisation approach, JOLs are influenced by three streams of cues, where the mnemonic stream is partially fed by both intrinsic and extrinsic streams.

This complex model seems to better explain disparities between JOLs and performance than the direct-access approach, because a biased judgment of any of the multiple variables on which JOLs are based may result in decreased accuracy. Therefore, the chance of a biased judgment is greater with a larger number of variables. One variable that has the potential to affect JOLs is feedback. Further, Koriat, Ma'ayan, and Nussinson (2006) explicitly argued that mnemonic cues are based on feedback on one's previous behaviour. It is this notion that formed the basis for the current study.

The abovementioned disparity between JOLs and performance can be investigated using the two most prominent measures for JOL accuracy: resolution and calibration (Lichtenstein & Fischhoff, 1977). Yaniv, Yates, and Smith (1991) explained resolution (or discrimination) as the ability to distinguish between two classes of items (e.g., items that will be recalled and those that will not). The larger the difference between JOLs for both class of items the better the resolution. In contrast, calibration measures the accuracy of probabilistic judgments (e.g., assessing the likelihood of future recall on a 0-100% scale) by calculating the correspondence between predictions and performance (Yaniv, Yates, & Smith, 1991). For example, in terms of the proportion of items recalled at each level of JOLs, of all items given 80 percent JOLs, how many were actually recalled. If the number of recalled items (for 80 percent JOLs) is below 80 percent this inaccuracy is called overconfidence, if

it is exactly 80 percent (i.e., the objective and subjective probabilities of recall match) calibration is perfect, and when recall is above 80 percent then calibration shows underconfidence. Although both calibration and resolution measure JOL accuracy these measures are differently affected by learning. According to Koriat (1997), repeated practice impairs calibration but it also improves resolution.

The impairment of JOL accuracy has been comprehensively researched in the context of the underconfidence with practice (UWP) effect. This phenomenon, described in detail by Koriat, Sheffer, and Ma'ayan (2002), can be observed when participants study the same items over a number of study-test phases. Calibration between participants' JOLs and performance tends to show overconfidence after the first learning cycle (i.e., JOLs overestimate actual recall performance), but underconfidence in the following cycles (i.e., JOLs underestimate actual recall performance). In the present study, we were interested in the effects of repeated learning on JOLs when learners receive assessments – settings typical for education.

The Interplay of Metacognition and Feedback in Education

The effects of feedback on metacognition have been extensively investigated in the context of education (Butler & Winne, 1995; Dunlosky & Metcalfe, 2008; Dunlosky & Rawson, 2012; Hattie & Timperley, 2007). These effects vary according to the type, timing, and source of feedback, as well as other conditions including, for example, the ability to interpret feedback (Hacker, Bol, Horgan, & Rakow, 2000). Thus, the ability to reflect on one's state of knowledge is a pivotal part of a successful learning process. There is also considerable research supporting the notion that feedback is the foundation of metacognitive beliefs (Butler, Karpicke, & Roediger III, 2008; Dunlosky & Metcalfe, 2008; Hattie & Timperley, 2007). Hence, the accuracy of metacognitive predictions is, at least partially, a function of accurate

feedback. This accuracy can also be affected by external conditions, for example, the timing of the judgment (Nelson & Dunlosky, 1991) and by cognitive variables, including the perceived difficulty of the information being studied (Thiede & Dunlosky, 1999). Ultimately, feedback plays a prominent role in forming metacognitive beliefs because it supports the monitoring of one's knowledge. Metacognition is, according to Nelson and Narens (1990), a dynamic interplay between monitoring of learning and control over learning. In this interplay the accuracy and distortions of monitoring affect the quality of learning. Further, Thiede, Anderson, and Therriault (2003) and Thiede and Dunlosky (1999) highlighted the importance of monitoring and argued that a learner can optimally allocate their resources only when the monitoring is accurate. Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) explicitly argued that, in the context of practice testing – a highly effective learning method - feedback is an important monitoring tool.

The effect of feedback on monitoring depends on the feedback type. For example, task-level feedback provides specific information about the performance of a particular task as opposed to general feedback, which provides summary information on student progress (Shute, 2008). Marks are a common example of general feedback, used at universities worldwide (Biggs, 1999; Sadler, 2005). This form of assessment feedback shows student performance on a task, often on a 100 point scale.

Shute (2008) reviewed the body of research on feedback and found that specific and timely feedback is generally more supportive of learning than general feedback, admitting however that the mechanism behind this phenomenon remains unclear. Kluger and DeNisi (1996) found a similar phenomenon in their meta-analysis, although they too could not elucidate its basis. More recently, Goodman,

Wood, and Chen (2011), in an experiment comparing the effects of different levels of feedback on project management tasks, found that the more specific the feedback, the better the outcome of the learning. However, the transfer of this learning to other tasks is impaired when the feedback is highly specific. Moreover, Goodman et al. (2011) argued that feedback has negative effects on explicit information processing, such as planning and assessing actions. Therefore, highly specific feedback improves learning of a feedback-related task but decreases the transfer of learning to other tasks due to the decreased level of cognitive engagement. Lam, DeRue, Karam, and Hollenbeck (2011) also challenged the generality of the beneficial effect of feedback. They found that whilst feedback can improve learning performance, it can also decrease performance when provided with such high frequency as to overwhelm the cognitive capabilities of participants.

According to Koriat (1997), feedback that relates specifically to item-by-item performance enhances the accuracy of metacognitive judgements. For example, feedback can alleviate the metacognitive distortion of the UWP effect. Koriat (1997) found that participants who received feedback in the first study-test cycle showed reduced underconfidence after subsequent study sessions when compared to participants who received no feedback. The feedback was presented in the form of a high-pitch sound when participants did not produce a correct answer. This effect was observable even though the intrinsic characteristics of the items (for example difficulty) and their extrinsic characteristics (for example study environment) were the same in both conditions. Koriat argued that feedback that provided additional indicators (i.e., mnemonic cues) enhanced calibration. These cues (in the form of feedback) signalled the probability of target recall to the participants. Therefore, Koriat's experiment supported the view that feedback can improve metacognition.

However, feedback did not improve recall. This can be explained by the controlled setting of the experiment which allowed for monitoring of the learning progress (via feedback) but did not allow the control of behaviour (e.g., increasing study time for more difficult items).

Different types of feedback, however, have varying effects on task performance and metacognition. The effect of item-by-item feedback on metacognition is complex and can be counterintuitive. Kornell and Rhodes (2013) showed that participants who received feedback after recalling a target word produced significantly less accurate JOLs than participants who did not receive feedback. The feedback was presented by showing the cue-target word-pair immediately after the recall. Therefore, this condition increased the exposure to the studied word-pairs, which was reflected in the significantly higher recall for this group. However, it seemed that participants did not fully acknowledge this additional study in their predictions because, in the group which received feedback, resolution was significantly lower than in the group with testing without feedback. Thus, in some cases, feedback does not improve metacognitive insight, which fits with the idea that participants do not recognise the value of additional learning opportunities in these multi-cycle paradigms.

In regards to the effect of feedback in applied settings (i.e., marks), Crooks (1988) argued that specific feedback has a significantly higher positive effect on a student's performance than general feedback. Butler's (1988) findings went even further – that grades have a null effect on future performance. Butler investigated the effect of feedback in the context of student motivation and argued that the key to learning is intrinsic, task-oriented motivation, reflected by student engagement, which in turn supports performance. Butler hypothesised that feedback in the form of

grades initiates ego-involved motivation focused on the outcome of the external assessment. Butler argued that this ego-involved motivation undermines performance because it shifts student engagement from learning to the outcome of learning. Butler (1988) compared engagement and performance between students who received marks, comments, and both. Only task-specific comments supported learning; students who received marks did not make learning progress. Butler also found that the provision of marks, when accompanied by specific feedback, cancelled out the positive effects of specific feedback. Butler argued that mixing specific feedback and marking is as unhelpful in learning as marking alone because when students are presented with both forms of feedback they tend to ignore specific comments and focus on the marks.

Other research suggested that the effect of summary feedback on metacognition is similarly negligible on metacognition as it is on performance (Hacker et al., 2000). For example, Hacker et al. (2000) compared the effect of prior performance and prior predictions on subsequent predictions. The study investigated predictions and outcomes of three educational psychology course exams taken consecutively by 99 students. Self-assessment skills were a major focus throughout the course. Additionally, the participants took practice multiple choice tests a week before the exams, which served as a basis to self-assess their strengths and weaknesses. Students received an answers key and textbook page numbers corresponding to the questions. The participants were requested to learn from their errors and discuss the content of the tasks with peers and the instructor. In the course of the actual exams, the participants made predictions about their marks before and after the multiple choice tests. After the first two exams, the students received their marks and were encouraged to use this feedback in their preparation for the third

exam. Hacker and colleagues hypothesised that students who had been trained to understand that the best predictor of future performance is their prior performance would increase their reliance on the feedback from past performance in making predictions about consecutive tests. Thus they would utilise the mnemonic cue in line with Koriat's model (1997). Contrary to the hypothesis, the students did not increase their reliance on past performance in making predictions but rather anchored their judgments on previous judgments. This result suggested that student metacognitive beliefs may be insensitive to summary performance feedback in the form of marks. Hacker et al. concluded that Koriat's (1997) findings, which support the effect of feedback on metacognition, cannot be generalised to more general types of feedback.

Conclusions as to the ineffectiveness of marks, drawn from Hacker et al. (2000) and Butler (1988), seem to undermine the beneficial effects of marks on future learning. However, it should be noted that this research was run in classroom settings, which has a strong applied value but which is also more open to confounds than controlled experimental settings. In an observational study the investigator's control over a predictor variable is limited, therefore it is possible that individual students have different learning experiences regarding, for example, instructor attention, studying time outside the classroom, classroom parameters (e.g., distance to visual aids), and interactions between students. Further, observational settings do not allow for the random allocation of participants to conditions and there is no control group, as was the case in Hacker et al.'s study (2000). Therefore, substantiating these strong claims about the negligible effect of marks on metacognition requires support from controlled experiments that can more clearly elucidate the mechanisms underlying the effects.

In the literature, the effects of various factors on JOLs have been investigated mainly for item-by-item predictions, elicited for each studied item (often a word-pair). Another category of judgments, global or aggregate JOLs, predict proportions of items successfully recalled from the entire list of studied items (Mazzoni & Nelson, 1995). Thus, global JOLs show the absolute number or percentage of items predicted to be recalled, whereas item-by-item JOLs represents the likelihood of a single item being recalled. The recent study of Geurten and Meulemans (2017), ran in experimental settings, supported the notion that even general feedback can improve metacognitive accuracy. In that study, young primary school children were required to remember the target word presented with a cue word. Children in one condition received accurate feedback on the number of correctly remembered word-pairs. Children in the other condition did not receive feedback. The experiment involved two study-recall phases with a different set of words in each phase. After the second study session, the children judged how many words they would remember. Children who received global feedback on their performance after the first study session had significantly better calibration than the children in the no feedback condition. This finding suggests that global feedback can improve global predictions of performance.

The effectiveness of general feedback in improving calibration can also be supported indirectly from the perspective of the anchoring-and-adjustment heuristic proposed by Tversky and Kahneman (1974). Tversky and Kahneman argued that people make assessments by adjusting their predictions from a point of reference. This psychological anchor, which initiates prediction, can be a product of the question formulation or a partial analysis of the task. For example, in Tversky and Kahneman's (1974) experiment, the researchers asked participants to assess the

percentage of African countries in the United Nations. The participants were allocated to conditions by the spinning of a wheel of fortune which indicated initial values of estimation. The participants then estimated first if this value was too high or too low, and then provided their own estimate. For the initial values of 10 and 65 percent, the mean estimation was 20 and 45 percent respectively. This finding revealed that participants anchored their estimations to the values provided, even though they knew that these values were provided randomly.

Scheck and Nelson (2005) provided evidence that the anchoring effect explains the underconfidence with practice (UWP) effect. They argued that people naturally anchor their JOLs at the range between 30 and 50 percent and that the JOLs shift towards this anchor when the actual likelihood of recall is greater than 50 percent. Geurten and Meulemans (2017) provided evidence that external general feedback can constitute a psychological anchor to which performance predictions are adjusted. They tested early primary school children in six conditions, feedback (or no feedback) and anchor (low, high, or no anchor), with eight students per cell, who were presented with both easy and difficult memory tasks. Children who were provided global feedback about their prior performance shifted their subsequent predictions towards the score received. The children adjusted their prediction to the feedback received regardless of the difficulty of their actual task. Thus, research on the effect of summary feedback on metacognition is far from conclusive and further investigation is needed. In our study, we were interested in the effect of marks, and the accuracy of feedback received, on student metacognition.

The Effects of Biased Feedback

While the importance of understanding the feedback effect (or its absence) is not under question, some aspects of feedback have been under-researched. Both

experimental and quasi-experimental research on feedback effects on metacognition has focused only on accurate feedback, despite the fact that inconsistency in student assessment has long been an important issue for educators and students alike (Henzi, Davis, Jasinevicius, & Hendricson, 2006; Sadler, 2005). Inaccurate feedback may have multiple bases. Apart from assessment differences between individuals (Birenbaum, 1997), there is evidence that feedback in the form of marks can be unconsciously biased (Rosenthal & Jacobson, 1968). In some situations assessors can also be consciously biased. For example, Dee, Dobbie, Jacob, and Rockoff (2016) claimed that assessors of New York's high school exit exams, inflated 40 percent of marks which were just below the grade threshold and that this inflation increased the probability of graduation for low achievers by 27 percent. In the Australian context, Steenkamp and Roberts (2017) argued that institutional pressure on the student pass rate coupled with limited resources, at an Australian university, drive assessors to achieve this goal by inflating marks. Crosby and Monin (2007) pointed out that Black students have a lower chance than White students to receive feedback warning them of potential academic difficulty. They found that Caucasians are strongly motivated to provide information without racial prejudice, which increases the likelihood that they withhold discouraging, but valid, information from Black students. There is also evidence of the phenomenon of feedback inflation within educational institutions across time (Rojstaczer & Healy, 2012; Vinton & Wilke, 2011), while Kuh and Hu (1999) also identified institutions where grade deflation took place over time. Despite considerable evidence of assessment distortion, the effect of inaccurate feedback on metacognition has not been thoroughly investigated. Inaccurate assessment is a serious issue because feedback distortion can diminish its

twofold educational goal: performance monitoring and learning facilitation (Wass, Van der Vleuten, Shatzer, & Jones, 2001).

It would be logical to assume that the effects of distorted feedback on metacognition would be a derivate of the effect of accurate feedback. Therefore, any investigation on biased feedback should be focused on the differences between the effects of accurate and deflated feedback and accurate and inflated feedback. It can be assumed that biased feedback would affect mnemonic cues (Koriat, 1997) similarly to accurate feedback, but that the strength of this effect would vary from the accurate feedback effect. The direction and the scale of this variation would be proportional to the direction and scale of the feedback bias. A similar mechanism could be justified by the anchoring-and-adjustment hypothesis (Tversky & Kahneman, 1975), where biased feedback would establish a different reference point than would accurate feedback. Therefore, the biased anchor could bias the adjustment from this anchor. Again, the scale and the direction of this adjustment would be a function of the biased feedback.

Basis of the UWP Effect

Koriat et al. (2002) argued that the UWP effect is a particularly robust phenomenon which is not reversed when manipulations are applied. These manipulations included the mode of study time allocation (fixed and self-paced), the varied difficulty of items (easy and difficult), and the presence or absence of an incentive. The UWP was maintained even when it was attenuated with feedback (Koriat, 1997). Therefore, the basis of the underconfidence did not derive from the underassessment of previous responses.

Koriat presented a list of possible explanations for the UWP mechanism, suggesting that the cue-utilisation model (1997) is amongst the most promising.

According to this hypothesis, in the course of repetitive study sessions, participants shift their basis of judgment from the task analysis to the mnemonic heuristic of past performance. This shift of focus towards past experience would explain why participants fail to fully analyse and acknowledge the effect of additional learning. Therefore, the progress in learning would not be reflected by a sufficient adjustment of JOLs.

Interestingly, Koriat et al. (2002) admitted to making a sampling error in the earlier study (1997) which argued that feedback, although not averting the UWP effect, did reduce the UWP effect in item-by-item JOLs. Nevertheless, Koriat did not repeat this experiment. The current study aimed to at least partially clarify the issue of the effect of feedback on JOLs, extending the question from item-by-item to global JOLs and investigating their sensitivity to feedback.

The Current Study

We examined the effect of accurate feedback, in the form of marks, against a no-feedback control group, on both item-by-item and global JOLs. We also examined this effect in relation to two variations of feedback: inflated and deflated against an accurate feedback group.

Participants studied 50 Swahili-English word-pairs in three study-test phases. Participants provided item-by-item JOLs after studying each word-pair and a global JOL (predicting the total percentage of words they would recall) after each study session. After the first and the second testing sessions, participants received feedback as per their condition. For the main hypothesis, we analysed the variables from phase 2 and phase 3 of the experiment because in phase 1 the outcome could not be affected by feedback. The dependent variables consisted of recall accuracy, item-by-item JOLs, global JOLs, resolution, the overconfidence/underconfidence scale and

the perception of feedback measure. The hypotheses involved general expectations related to well-grounded research on the UWP effect and novel hypotheses related to the effects of feedback.

Recall performance. In line with Nelson and Dunlosky (1994), it was expected that recall performance would increase with every study session. We tested for any effects of the feedback manipulation on recall performance. However, we did not expect that the performance between conditions would be significantly different because, even if feedback affected monitoring, the fixed learning conditions did not allow for improvement through control over the learning process.

Resolution. In regards to resolution (i.e., an index of metacognitive monitoring), we expected that participants' item-by-item JOLs would discriminate between remembered and not remembered items (Ariel & Dunlosky, 2011). Resolution – the predictive ability to discriminate between correct and incorrect responses – was measured with ANDI – the Adjusted Normalized Discrimination Index (Yaniv, Yates, & Smith, 1991).

The UWP effect. We expected that the experiment would replicate well-grounded findings of the robustness of the UWP effect (Koriat et al., 2002) which, researchers have argued, is too strong to be fully reversed by the manipulation of only one factor (feedback) of the many factors which are potential bases for JOLs. Therefore, we hypothesised that calibration in phase 2 and in phase 3 would show underconfidence for both global and item-by-item JOLs in all conditions.

Effects of feedback. ***The hypothesis regarding the effect of accurate feedback.*** In regards to the effect of feedback, we first examined whether feedback in the form of marks improves the calibration of item-by-item and global JOLs. Koriat's (1997) cue-utilisation model supports this hypothesis by suggesting that

feedback feeds into the mnemonic cue, which is in part the basis for JOL. The notion that feedback improves JOL accuracy can also be derived from the anchoring-and-adjustment heuristic (Tversky & Kahneman, 1975), which suggests that feedback can serve as a starting reference point for making predictions. Empirical support for the beneficial effect of feedback on JOLs (Koriat, 1997) is, however, problematic given Koriat et al.'s (2002) admission of a possible sampling error in Koriat's (1997) study. Also, Hacker et al. (2000) questioned the generalisation of Koriat's (1997) findings on the effect of item-by-item feedback to the effect of marks on subsequent judgments. On the other hand, however, Geurten and Meulemans (2017) provided evidence that the positive effects of feedback on JOLs can be generalised to the effect of marks (scores) on subsequent global predictions.

From the above literature review, there are theoretical arguments to support the effects of marks on JOLs and, in consequence, on calibration. There is also some empirical support for the feedback effects, although it is also possible that some of these results were confounded (Koriat, 1997). We argue that claims that marks have a null effect on metacognition are not sufficiently supported and these claims need to be investigated in more controlled settings than have been done to date (for example by Hacker et al., 2000). Therefore, we cautiously hypothesised that the provision of accurate marks would improve the calibration of item-by-item and global JOLs when compared with the no feedback condition. This improvement in calibration was operationalised in terms of changes in the overconfidence/underconfidence (OU) index (Lichtenstein, Fischhoff, & Phillips, 1982). The scale of the OU index ranges from *1* to *-1*, with *0* representing perfect calibration.

Hypotheses regarding the effect of inflated and deflated feedback. We extended our investigation of the effects of accurate feedback to inflated and deflated

feedback. Specifically, we hypothesised that inflated feedback can increase item-by-item and global JOL calibration and that deflated feedback can decrease both types of JOL accuracy when compared with the effect of accurate feedback on JOLs. This effect is also operationalised in terms of the OU index. Koriat (1997) argued that performance on earlier tasks is a valid predictor of JOLs. Therefore, we proposed an extension of Koriat's cue-utilisation hypothesis to argue that biased feedback in the form of marks would predict biased JOLs. Thus, we assumed that the direction of biased marks (inflated, deflated) would predict the direction of bias for JOL accuracy. This hypothesis was also supported by the potential anchoring effect of feedback (Geurten & Meulemans, 2017).

We expected the manipulation to produce similar effects for both item-by-item and global JOLs because the basis of both types of JOLs can be explained by the same cue-utilisation model (Koriat, 1997). However, within the same model, it is also possible that item-by-item and global JOLs vary in sensitivity to marks. It could be speculated that general feedback in the forms of marks would also create theory-based cues (related to general belief about one's skills), aside from experience-based cues (related to learning of particular items). Therefore, it is possible that global JOLs, which produce general judgment, would utilise general feedback (theory-based cue) more than item-by-item JOLs. Hence, the effect of marks may be more pronounced for global JOLs than for item-by-item JOLs.

The final examination in the study was a perception of feedback test which investigated the extent of alignment between a participant's expectations of marks and the actual marks received. This measure was designed to provide additional insight into the study outcomes. This hypothesis is exploratory in nature and,

therefore, we looked for significant differences between conditions. The strength of differences between conditions, however, was difficult to anticipate.

Method

Participants and Design

The participants comprised 80 adults (31 males, M_{age} = 29.61 years, *SD* = 12.76). Participants received either $15 compensation for their time or one-hour research participation credit (1st-year Psychology students at UTAS). A reward of $200 was drawn from amongst the best 10 performers to facilitate recruitment and participant engagement. Participants were recruited by email, from the Psychology Research Participants Pool, or from advertisements around campus. The sample comprised domestic and international students. The only exclusion criterion was experience learning Swahili (because the task involved learning Swahili-English word-pairs). No participants were excluded based on this criterion. The data of one participant were excluded for non-engagement in the task (the same two non-task related words were used in most of the 3x50 recall trials, and all JOLs were at 50 percent). Another participant was not able to finalise the task due to IT complications. Participants were randomly assigned to four conditions with 20 participants per cell.

The study used a 4 x 2 (3) mixed factorial design. The between-subject factor was the type of feedback received after the first and the second testing sessions: no feedback (1); accurate feedback (2); feedback inflated by 32 percent of the actual score (3); feedback deflated by 32 percent of the actual score (4). The within-subjects factors were phases of learning and testing (2 and 3).

The dependent variables included: item-by-item JOL magnitude (on a scale of 0 to 100) and global JOL magnitude (on a scale of 0 to 100); recall performance;

calibration between JOLs and performance expressed in the directional OU index (Lichtenstein et al., 1982). On this scale, ranging from *-1* to *1*, *0* represents a perfect match between recall and JOLs. Values above *0* indicate overconfidence and values below *0* indicate underconfidence. The OU measure for item-by-item JOLs was calculated as a mean of the OU values at each level of JOL. For global JOLs the OU measure reflects the averaged ratio of global JOLs and recall. Resolution was expressed in ANDI – the Adjusted Normalized Discrimination Index (Yaniv et al., 1991) which ranges between 0 (no discrimination) and 1 (perfect discrimination). A perception of feedback check was measured on a scale from 1 (feedback was lower than predictions) to 10 (feedback was higher than predictions) where 5 represented a match between predictions and feedback.

Materials

The entire experiment was run on a computer with software prepared by the Software Engineer for the Tasmanian Cognition Laboratory. The software used Jaro–Winkler distance (Winkler, 1990) to detect and accept answers with simple spelling errors. The experiment used 55 items (see Appendix 1) from a pool of 100 Swahili-English word-pairs developed by Nelson and Dunlosky (1994). One item was used for a practice session and four items were used as primacy buffers and excluded from the analysis. The remaining 50 word-pairs were selected for learning and testing sessions. According to norms developed by Nelson and Dunlosky the mean recall for the pool of the 100 word-pairs was 0.14 (*SD* = 0.10) in phase 1, 0.42 (*SD* = 0.16) in phase 2, and 0.63 (*SD* = 0.14) in phase 3. The mean recall for the 50 selected items was .20, .50, .69 respectively. *SD*s for the selection was not available, therefore we used the pool *SD* from phase 2 as a reference point for feedback manipulation.

Procedure

The procedure followed that applied by Nelson and Dunlosky (1994). As the participants arrived they read the information sheet and signed a consent form. All instructions regarding the experiment were displayed on a computer screen. A one-item practice session preceded the experiment. The researcher was available for any enquires throughout the experiment. The experiment, in all conditions, involved three cycles. Every cycle comprised a learning phase followed by a testing phase. The learning phase involved 54 English-Swahili word-pairs which were displayed on a computer screen for 10 seconds each. The first four word-pairs were a primacy buffer and were not included in the practice session. Participants provided item-by-item JOLs after studying each word-pair and predicted the total percentage of words they would recall after each study session (global JOL). In the testing phase, the participants were presented with a Swahili cue word, after which they typed the target word and clicked *continue*. Information was displayed on the computer screen indicating that the participant had 15 seconds to accomplish the task. When the time expired another cue word appeared automatically on the screen. At the end of the experiment, as a perception of feedback check, the participants assessed whether the feedback received matched their predictions.

Results

General Findings

Before addressing the main question regarding the effect of feedback on underconfidence for item-by-item and global JOLs, more general findings are reported.

Data Screening. The design of the study prevented the possibility of missing data. All data were screened for outliers, with some mild outliers identified and

removed. The data pattern did not change following outlier removal, therefore the original values were reported, with the exception for ANDI data, where outliers were permanently removed. The normality of the data was inspected with the Kolmogorov-Smirnov and Shapiro-Wilk tests in conjunction with Q-Q plots, histograms and skewness and kurtosis measures as advised by Field (2013). The inspection revealed that the data were skewed, and a square root transformation and a reverse score transformation was applied to correct this skew. However, as the pattern of results based on the transformed data was identical to that for the original data, raw data are reported. The highest values of the skew were related to phase 1 of the experiment which was not tested in regards to the main hypotheses, therefore it was assumed that the robustness of ANOVA (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010) and the relatively low skew values would be sufficient to deliver reliable results for the main analysis. Levene's tests in all but one (remediated with the Brown-Forsythe test) of the following analyses revealed linearity of the data. A single violation of data homoscedasticity was corrected with the Greenhouse-Geisser correction. Both corrections were noted below.

Demographics. Two one-way ANOVAs revealed that there were no age differences between groups: F (3, 76) = 0.80, p = .497, η_p^2 = .03, and that males and females were similarly distributed F (3, 76) = 0.46, p = .712, η_p^2 = .02.

Recall performance. We tested for any effects of the feedback manipulation on recall performance. A two-way 4 (feedback condition) x 3 (phase) mixed ANOVA, with Greenhouse-Geisser correction, revealed that there was no significant main effect of condition on the recall performance, F (3, 76) = 0.37, p = .773, η_p^2 =.02. The interaction of condition and study phase was also non-significant, F (4.85, 122.73) = 0.56, p = .727, η_p^2 = .02. In line with our prediction, the main effect of

study phase on recall performance was significant, F (1.62, 122.73) = 715.01, p < .001, η_p^2 = .90. Two within-subjects t-tests with a Bonferroni correction (to p < .025) revealed a significant increase in recall between phase 1 (M = .24, SD = .17) and phase 2 (M = .54, SD = .24), t (79) = 23.56, p < .001, 95% CI [.28, .33], d = 2.34, and between phase 2 and phase 3 (M = .71, SD = .22), t (79) = 18.17, p < .001, 95% CI [.15, .19], d = 1.45.

Resolution. We hypothesised that participants would be able to discriminate between items that they would and would not remember (i.e., based on item-by-item JOLs). However, we made no prediction about the effect of the feedback manipulation on resolution. JOLs predicted recall in all phases, as one-sample t-test revealed that overall ANDI in phase 1 (M = .19, SD = .20) was significantly above zero, t (79) = 8.54, p < .001, 95% CI [.14, .23], with a large effect size d = 0.95. Thus, in line with the hypothesis, the participants were able to discriminate between items they would and would not remember. A 4 (feedback condition) x 2 (phase 2 and 3) mixed ANOVA, revealed a non-significant main effect of condition on ANDI, F (3, 72) = 0.26, p = .853, η_p^2 = .01. Further, the interaction of condition and study phase for global JOLs was non-significant, F (3, 72) = 1.98, p = .125, η_p^2 = .08. Thus, there was no evidence that the feedback manipulation affected resolution. Overall ANDI between phase 2 (M = .21, SD = .17) and phase 3 (M = .29, SD = .23) increased significantly, F (3, 72) = 9.35, p = .003, η_p^2 = .12.

The UWP effect. We expected to see the UWP effect for both item-by-item and global JOLs. An inspection of calibration graphs (Figure 1) for item-by-item JOLs (plotting variations in recall as a function of JOL magnitude) reveals the typical UWP effect for all conditions. Participants typically showed overconfidence in phase 1 and underconfidence in phases 2 and 3. The UWP effect was also observed on the

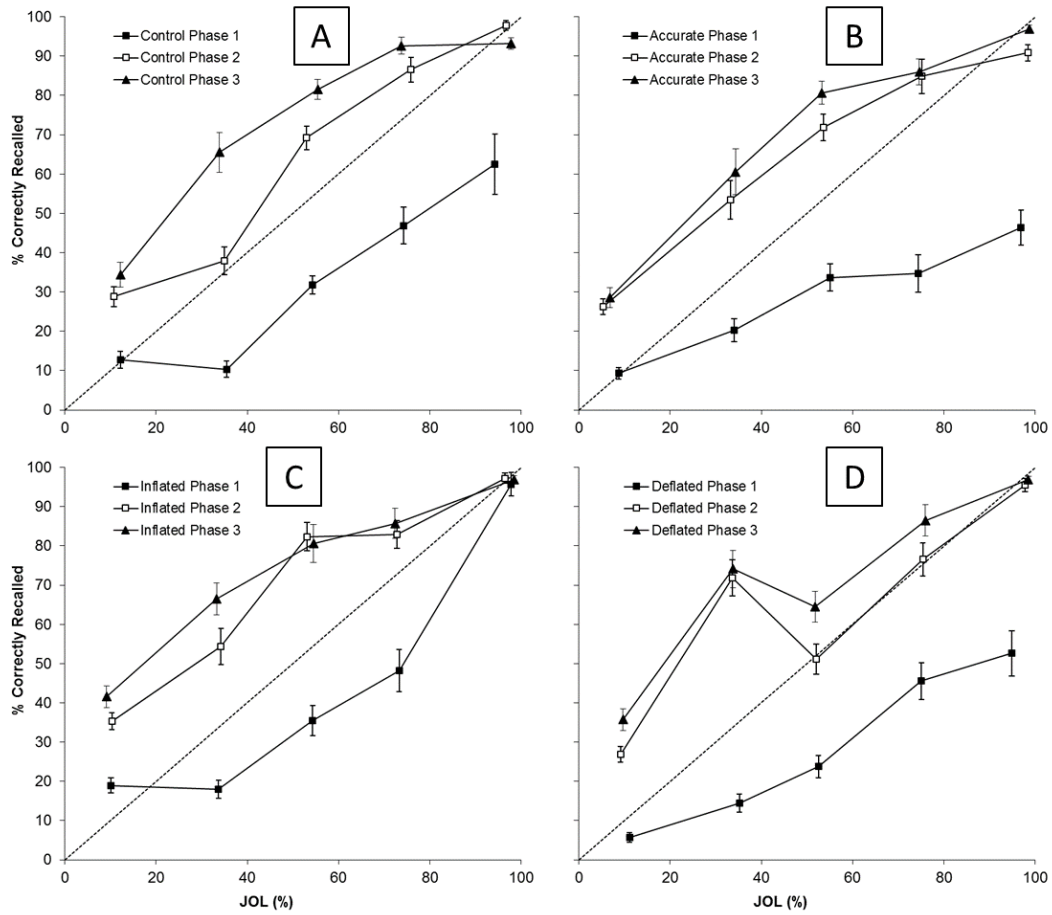

Figure 1. Calibration curves between the correctly recalled items and the item-by-item JOLs for: no feedback (A); accurate feedback (B); inflated feedback (C); and deflated feedback (D) groups. The error bars show standard errors. The diagonal line represents perfect calibration. Data placed below the diagonal line show overconfidence, or underconfidence when located above the diagonal line.

OU index for global JOLs. A within-samples t-test revealed that, in line with the hypothesis, in the first phase of the experiment, when all the groups received the same treatment, global JOLs (M = 36.50, SD = 18.56) were significantly higher than recall accuracy (M = 23.93, SD = 16.72), t (79) = 6.31, p < .001, 95% CI [8.61, 16.53], d = 1.42, indicating overconfidence. In phases 2 and 3 (i.e., following

exposure to the feedback manipulation), mean recall exceeded mean global JOLs (see Figure 3), indicating underconfidence. However, underconfidence was not significant for accurate and inflated feedback as 95% CI for global JOLs and recall accuracy overlapped for these conditions. Therefore, UPW effect for global JOLs was significant only for deflated feedback and partially (no effect in phase 2) for the accurate feedback condition.

Effects of Feedback

We hypothesised that accurate feedback would increase the accuracy of item-by-item and global JOLs, compared to the no feedback group. We also expected that inflated feedback would moderate underconfidence more than accurate feedback for item-by-item and global JOLs, because higher/lower feedback would potentially provide a higher/lower reference point which would be used as a cue in probability predictions. Finally, we expected that deflated feedback would increase underconfidence when compared with the effect of accurate feedback for both types of JOLs. We report the investigation of these hypotheses separately for item-by-item and global JOLs. The effects of feedback analysis involved only phase 2 and phase 3 of the experiment, as the phase 1 results reflect performance prior to the feedback manipulation.

Item-by-item JOLs. A 4 (feedback condition) x 2 (phase 2 and 3) mixed ANOVA revealed, contrary to the hypothesis, that there was no significant main effect of condition on item-by-item JOL magnitude, F (3, 76) = 0.22, p = .886, η_p^2 = .01. The interaction between condition and study phase for item-by-item JOLs was also non-significant, F (3, 76) = 0.46, p = .710, η_p^2 = .02. The main effect of the study phase on item-by-item JOLs was significant, F (3, 76) = 159.61, p < .001, η_p^2 =

.68. The pattern presented in Figure 2 revealed that the mean item-by-item JOL increased with every study phase in all conditions, as did recall accuracy.

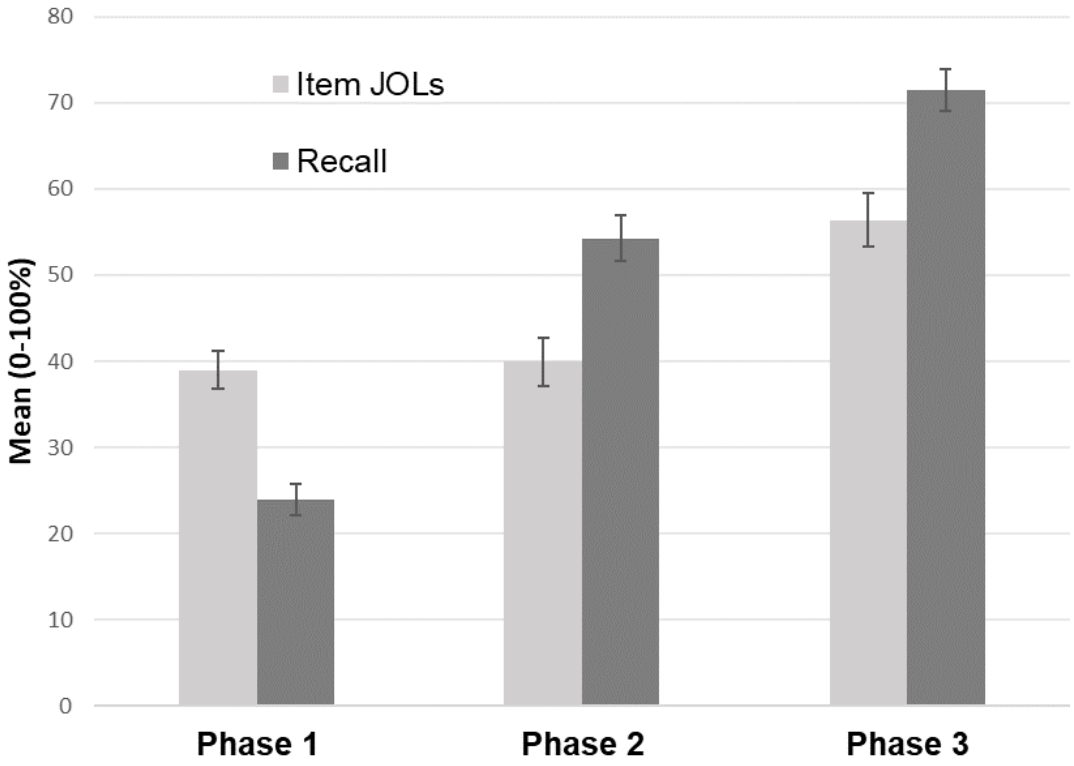

Figure 2. Overall magnitudes for item-by-item JOLs and accurate recall. The error bars show standard errors.

The OU index for item-by-item JOLs. A two-way 4 (feedback condition) x 2 (phase) mixed ANOVA revealed that, contrary to the hypothesis, the main effect of feedback on the level of underconfidence was non-significant, F (3, 76) = 0.61, p = .609, η_p^2 = .02, for item-by-item JOLs. Also, the interaction of phase and condition was not significant F (3, 76) = 0.58, p = .632, η_p^2 = .02. The effect of phase was also not significant, F (1, 76) = 0.18, p = .672, η_p^2 < .01. Thus, although participants typically shifted from over to underconfidence from phase 1 to phase 2, there was no evidence of any difference between phase 2 and phase 3.

Global JOLs. A two-way 4 (feedback condition) x 2 (phase) mixed ANOVA revealed, contrary to the hypothesis, that the main effect of condition on global JOLs was not significant, F (3, 76) = 1.89, p = .139, η_p^2 = .07. However, the interaction of condition and study phase for global JOLs was significant, F (3, 76) = 3.36, p = .023, η_p^2 = .12. A follow-up test of simple effects revealed that there was a significant main effect of condition, F (1, 76) = 1.47, p = .035, in phase 2 (despite not being detected by the mixed ANOVA), but not in phase 3, F (1, 76) = 1.47, p = .231. A one-way ANOVA for phase 2 confirmed a significant main effect of condition F (1, 76) = 3.01, p = .035, η_p^2 = .11. A post hoc Tukey test revealed that global JOLs for deflated feedback (M = 24.00, SD = 13.92) were lower (p = .023) than global JOLs for the no feedback condition (M = 44.00, SD = 23.71). However, we treated this result with caution because Levene's test for the ANOVA was significant F (3, 76) = 2.75, p = .049. Therefore, the Brown-Forsythe test was used, which confirmed the significant main effect of condition in phase 2, F (3, 68.20) = 3.01, p = .036. We followed with an independent t-test which, in line with the hypothesis, supported the Tukey test results that global JOLs for deflated feedback were significantly lower than for the no feedback condition t (38) = 3.25, p = .002, 95% CI [7.56, 32.45], d = 1.03, in phase 2.

The main effect of the study phase on global JOLs was significant, F (3, 76) = 69.74, p < .001, η_p^2 = .48. An inspection of the graphs in Figure 3 revealed that mean global JOLs increased significantly between phase 2 and phase 3 for accurate and inflated conditions (moderate overlap of 95% CI) and deflated conditions (no overlap). However, this increase seemed not significant for the no feedback condition as 95% CI overlapped more than moderately for this condition (Field, 2013).

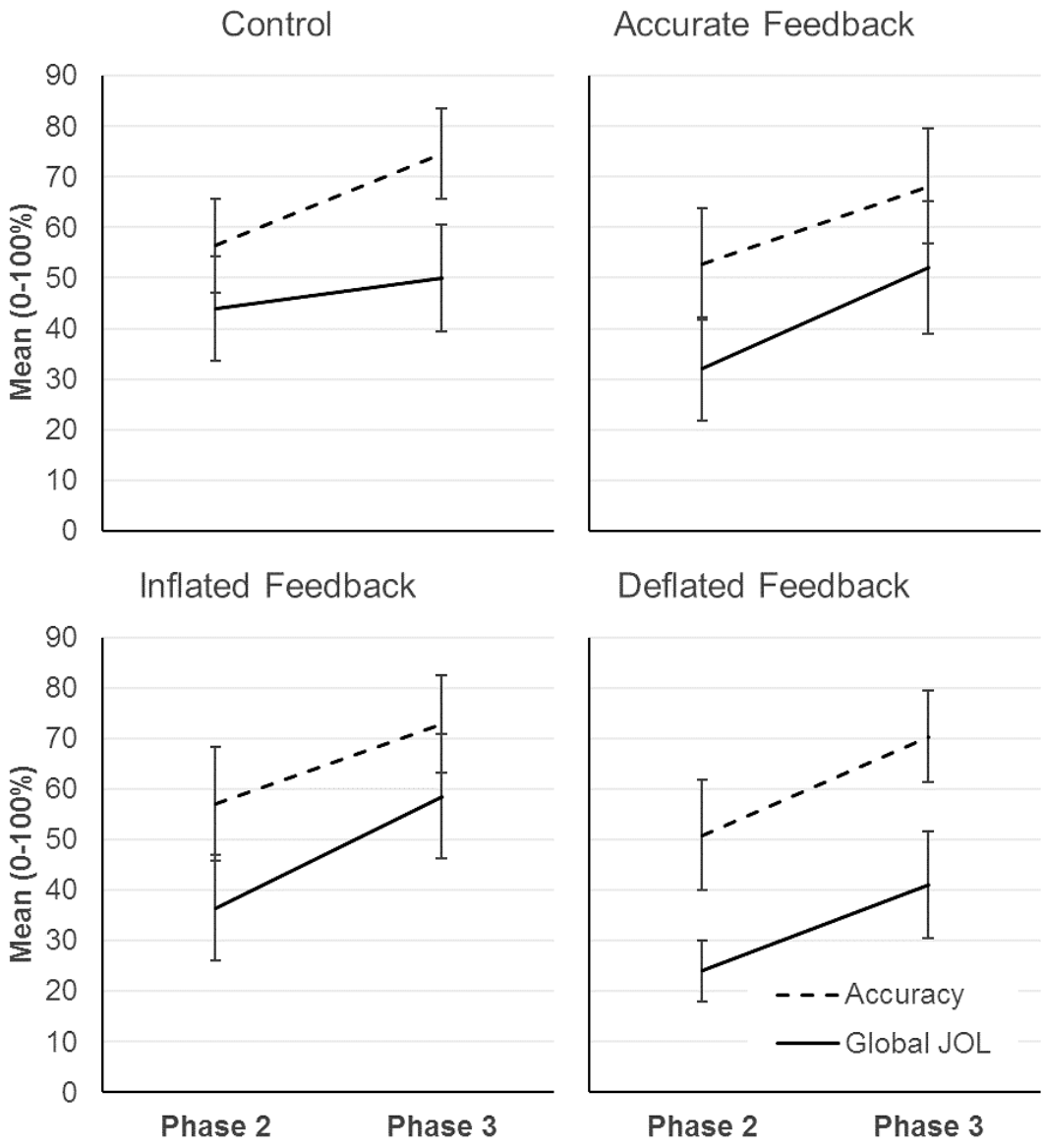

Figure 3. Comparison of the means of correctly recalled items and of the means of global JOLs. The error bars show 95% CIs.

The OU index for global JOLs. A two-way 4 (feedback condition) x 2 (phase) mixed ANOVA revealed that the main effect of feedback on the level of underconfidence was significant F (3, 76) = 4.49, p = .006, η_p^2 = .15, suggesting that the provision of feedback made a difference in the OU index for at least one of the four levels of feedback The main effect of phase on the level of underconfidence for

global JOLs was not significant F (1, 76) = 0.16, p = .692, η_p^2 < .01. This effect revealed that the OU index was not influenced by repeated learning. The interaction of phase and condition was significant F (3, 76) = 3.23, p = .027, η_p^2 = .11. This interaction effect indicated that the effects of feedback on the OU Index, between four levels of feedback, were different in phase 2 and in phase 3. To break down this interaction we ran a follow-up test of simple effects of condition within two levels of phase which revealed that there was a significant difference between groups in phase 2, F (3.76) = 2.92, p = .039, and also in phase 3, F (3.76) = 5.03, p = .003.

Effect of accurate feedback. A follow-up series of six independent-samples t-tests (see Table 1) was conducted to test the hypothesised effects of accurate feedback (compared to no feedback, inflated feedback, or deflated feedback) on underconfidence. In phase 2, contrary to the hypothesis, there was no significant difference between the group which received accurate feedback and the group without feedback. In phase 3, in line with the hypothesis, a similar medium effect size was observed in the reversed direction. Although these comparisons returned non-significant results, the effect sizes for differences exceeded the cut-off for a medium effect. These comparisons may, therefore, suffer from a lack of power. The effect sizes may suggest a genuine difference, but need to be interpreted with caution.

Effect of biased feedback. A cycle of t-tests (see Table 1) revealed, contrary to the hypothesis, that in phase 2 the levels of underconfidence were nearly identical for accurate and inflated feedback, whereas in phase 3 the difference was not significant and the effect size was trivial. Contrary to the hypothesis, the difference between accurate feedback and deflated feedback in phase 2 was not significant, but

Table 1

Effect of Feedback on the OU Index for Global JOLs

Pairwise Comparisons	No Feedback *M* (*SD*)	Accurate Feedback *M* (*SD*)	Inflated Feedback *M* (*SD*)	Deflated Feedback *M* (*SD*)	*t* (38)	*p*	95% CI	*d*
Phase 2								
No Feedback – Accurate Feedback	-.12 (.19)	-.21 (.12)			1.68	.101	[-.02; .19]	0.53
Accurate Feedback – Inflated Feedback		-.21 (.12)	-.21 (.16)		0.05	.964	[-.09; .09]	0.01
Accurate Feedback – Deflated Feedback		-.21 (.12)		-.26 (.15)	1.44	.159	[-.03; .15]	0.38
Phase 3								
No Feedback – Accurate Feedback	-.25 (.17)	-.16 (.12)			1.77	.085	[-.18; .01]	0.56
Accurate Feedback – Inflated Feedback		-.16 (.12)	-.14 (.16)		0.42	.675	[-.11; .07]	0.13
Accurate Feedback – Deflated Feedback		-.16 (.12)		-.29 (.10)	3.68	.001*	[.06; .20]	1.16

Note. For all groups n =20. A Bonferroni correction (.05/6) was used to establish t-tests significance level at $p < .008$*.

the effect size was medium, indicating some possibility that deflated feedback deepened underconfidence. In phase 3, in line with the hypothesis, deflated feedback significantly increased underconfidence, when compared with the accurate feedback condition. The effect size of this difference was large. An inspection of the graph (see Figure 3) comparing the differences between recall accuracy and global JOLs seemed to suggest that the significant effect of deflated feedback on underconfidence was rather due to lowered global JOLs than to reduced recall.

Perception of feedback. A one-way between groups ANOVA revealed a significant main effect, F (2, 57) = 4.86, p = .011, η_p^2 = .15, of condition on perception of feedback. Follow up one-sample t-tests, with the Bonferroni corrected significance level of $p < .013$, revealed that for participants who received deflated feedback (M = 5.05, SD = 1.67) this feedback matched their expectations because the difference from the matched expectation point of 5 was not significant, t (19) = 0.13, p = .895, d = 0.03, 95% CI [-.73, .83] and the effect size was trivial. However, participants in both the accurate (M = 6.70, SD = 2.11) and inflated feedback (M = 6.80, SD = 2.17) conditions indicated that the feedback received was significantly higher than the match point, t (19) = 3.61, p = .002, d = 0.81, 95% CI [.71, 2.69] and t (19) = 3.72, p = .001, d = 0.83, 95% CI [.79, 2.81], for the accurate and inflated feedback conditions, respectively. An independent samples t-test revealed that there was no significant difference between the accurate and inflated feedback conditions, t (38) = 0.15, p = .883, 95% CI [-1.47, 1.28], d = 0.05, therefore their perception of feedback was equally higher than expected.

Discussion

The current study investigated the effects of accurate and inaccurate feedback, in the form of marks, on item-by-item and global JOLs. The first main

hypothesis of the current study concerned the effect of feedback in the form of marks on the accuracy (OU) of item-by-item and global JOLs. We hypothesised that providing marks would increase the accuracy (OU) of both types of JOLs. The results do not fully support this hypothesis. However, medium effect sizes suggest that the expected effect could be significant with increased power. Secondly, we expected that inflated feedback would moderate underconfidence and that deflated feedback would aggravate underconfidence when compared with accurate feedback. The results partially support the second hypothesis. Interestingly, the results show divergent patterns for item-for-item JOLs and global JOLs.

Effects of Marks

The results reveal that the provision of marks, whether accurate, deflated or inflated, did not affect participants' recall performance. Feedback also did not affect resolution. Neither did we find evidence that participants incorporated feedback into item-by-item JOLs. The indifference of marks on these results was mirrored by the lack of feedback effects on levels of underconfidence. Therefore, contrary to the hypothesis, these results do not support the notion that the effects of concrete feedback (Koriat, 1997) can be generalised to the effects of marks. The failure of this generalisation is in line with Hacker et al. (2000). Thus, marks do not seem to improve student metacognitive assessment. The participants also discounted accurate and biased feedback in making their item-by-item JOLs. Therefore the expectation that biased marks feed into mnemonic cues as an experience related heuristic (Koriat, 1997) is not supported for item-by-item JOLs.

The effects of marks on global JOL accuracy is more complex. The differences between the accurate feedback and no feedback conditions were non-significant. However, medium effect sizes for these effects were observed in phase 2

and phase 3. Therefore, with a larger sample size, increased power would increase the chance of observing significant effects. The direction of the observed effect size in phase 2 is puzzling because it indicates that the accurate feedback condition generates higher underconfidence than the no feedback condition. However, in phase 3 the direction of this effect was the opposite and in line with the hypothesis. These inconclusive results did not fully support the findings of Geurten and Meulemans (2017) that global feedback improves calibration for global JOLs.

The most interesting results emerged in the effects of biased marks on global JOLs. Whereas students treated inflated feedback in a similar manner as accurate feedback in phase 2 and 3 (i.e., showing reduced underconfidence), deflated feedback significantly increased underconfidence when compared with accurate feedback in phase 3. This effect size was large, providing strong evidence that the negative bias in assessment increased underconfidence for global JOLs. In the previous phase (2) this difference was non-significant, however, the medium effect size suggested that with a larger sample size deflated feedback may generate significantly higher underconfidence than accurate feedback. The unique sensitivity of the participants to deflated feedback suggested that negatively biased marks compounded participants' general tendency to undervalue learning across trials (i.e., the mechanism typically thought to underlie the UWP effect). This effect is at least partially consistent with the cue-utilisation hypothesis (Koriat, 1997). Interestingly, accurate and inflated feedback seemed to neutralise the UWP effect for global JOLs, in contrast to the findings on the robustness of the UWP effect (Koriat et al., 2002).

The different effect of negatively biased marks on global JOLs, in contrast to the effects of accurate and positively biased marks, was also supported by the perception of feedback check. This measure indicates that participants perceived

deflated feedback as almost ideally matching their expectations, whereas the accurate and inflated feedback considerably exceeded expectations. The finding that deflated feedback perfectly matched the assessment expectations suggested that negatively biased marks were seen as fully credible, whereas both inflated and deflated marks were seen equally, as largely overestimated. The perception of feedback check was a novel way to measure a third-order cognition and supported the outcome of the metacognitive predictions. The participants' acceptance of negatively biased feedback as an expected outcome provides further insight into the UWP effect suggesting fundamental underconfidence in repeated learning tasks. It seems that people not only tend to discount the effects of learning, but they are more willing to accept information confirming their biased assessment than more realistic feedback.

The difference between effects of marks for item-by-item JOLs (no effects) and global JOLs (mixed effects) suggests that the basis for these two types of judgments is not the same. These different effects may be explained by a potential difference in the relative contribution of theory-based cues and experience-based cues to two types of JOL. According to Koriat (1997; 2007) theory-based cues generate judgements which are based on one's beliefs, for example in one's learning capacities. Koriat further argued that experienced-based cues influence judgments related to the learning experience, for example how difficult items are. The results of this study suggest that it is possible that marks have a stronger influence on theory-based cues (i.e.: beliefs in one's capabilities) than on experienced-based cues (i.e.: perception of learning efficiency). This may be because marks provide a summary score which is not related to a concrete item. Such summary feedback can be understood as the assessment of general skills, not as mere learning feedback. Therefore if global JOLs are more strongly based on theory-based cues than item-by-

item JOLs, and if marks feed primarily these theory-based cues, marks would yield stronger effects on global JOLs than item-by-item JOLs. Conversely, if item-by-item JOLs are more strongly affected by experienced-based cues, the relative weakness of these cues, in the provision of marks, would explain the indifference of participants to marks in JOLs generated for a concrete item.

The availability heuristic (Tversky & Kahneman, 1973) also suggests that the effect of marks would be more pronounced for global JOLs than for item-by-item JOLs, because when formulating global JOLs the memory of past marks seems more available than the sum of likely remembered items, whereas for item-for-item JOLs the memory of the last item seems more available than past marks.

This study successfully replicated other well-grounded phenomena related to JOLs: increased performance and JOLs with study phase; and partially, the UWP effect. Therefore, the conclusions derived from the novel hypotheses are additionally supported by the reliability of the overall design, which provided expected results in a number of key domains.

Implications and Follow-Up Studies

One implication of the study to educational practice is that marks alone are not helpful in the learning of particular items. For example, the lack of feedback effect on item level JOLs could suggest that if students receive only marks or grades after a test or exam, this information would not contribute to their metacognitive insight when learning particular information for a subsequent test or exam. Therefore, teachers and tutors need to rely on other ways of communicating progress to improve student metacognition. Follow-up studies could investigate the effects of other, more concrete and comprehensive types of feedback, to broaden the knowledge in the area already explored for example by Butler et al. (2008). It should

be remembered, however, that such generalisations of experimental effects to more naturalistic settings have limited validity (Cassidy, 2013).

Marks have some potential to affect students' global perceptions of their knowledge and/or ability. The significance and the direction of the effect of accurate marks on these global self-assessments have not been fully clarified in the present study; therefore, further studies with larger sample sizes would be needed. Such investigation has the potential to provide additional insight into the mechanism of academic self-efficacy – belief in one's academic capacity (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), as is discussed below.

The current study, however, showed strong and clear effects of deflated marks on general predictions. This may suggest that students pay particular attention to lowered marks and, mistakenly, they treat this understated assessment as more valid than accurate feedback. The implication of this finding may be that deflated marks affect student general self-assessment more profoundly than other types of feedback. This raises the question whether the effects of negatively biased marks on global JOLs can be generalised to negative effects on academic self-efficacy. Direct testing of this assumptions is not ethically acceptable because it would require the provision of biased academic results. Therefore, it could be only indirectly implied that negatively biased feedback may lead to lowered academic self-efficacy. This implication can be grounded in the finding that in the course of the semester academic self-efficacy decreases for students who receive low marks and it increases for high achievers (Zusho, Pintrich, & Coppola, 2003). If this assumption is correct, it would mean that deflated feedback could lead to negative effects of low academic self-efficacy such as lowered levels of goal settings and, in consequence, lowered engagement in learning (Zimmerman, Bandura, & Martinez-Pons, 1992). Also, the

studies investigating theories of intelligence has found that students with low levels of self-efficacy believe that intelligence is innate – that it cannot be improved by studying – and that such students cannot effectively control their studying (Komarraju & Nadler, 2013). Thus, it could be implied that negatively biased feedback, through its negative effect on self-efficacy, may trigger a self-fulfilling prophecy – a Pygmalion effect in reverse (Rosenthal & Jacobson, 1968).

Limitations

The current study is not free from limitations, the most important being insufficient power. It would seem that the sample size used, the power determinant, is sufficient for the replication of strong effects such as the UWP effect (Koriat et al., 2002). However, the hypothesis was related to much more subtle inferences – the difference in the UWP effect strength between groups. Therefore the group sample should have been significantly larger as the sought effect was more subtle (Rosenthal & Rosnow, 1991). Also, the hypothesised effects involved between-groups comparisons which in itself requires higher power than within-subject designs (Maxwell, Kelley, & Rausch, 2008). The power limitations did not allow for more detailed comparisons, which would require splitting the number of participants per group in half, for example, for high and low performers, and for easy and difficult items. Such comparisons would be possible in future studies with increased power.

Conclusion

The current study investigated the effects of accurate and biased marks on item-by-item and global JOLs. Although students did not seem to incorporate marks in their item-by-item JOLs, this summary feedback has some potential to affect global perceptions of ability. It seems that students are particularly vulnerable to deflated feedback, which makes them assess their general recall abilities as much

lower than students who received accurate or inflated feedback. This finding implies that students who receive negatively biased marks may have a lowered assessment of their skills, which can potentially lead to lower motivation and decreased engagement with learning. This finding highlights the possibility that overly harsh testing standards can be counterproductive. The effect of inflated and accurate marks on global JOLs has not been fully clarified and the emerging effect sizes encourage future studies in this direction. Such studies should have larger sample sizes to increase power.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39*, 171-184. https://doi.org/10.3758/s13421-010-0002-y

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development, 67*, 1206-1222. doi:10.1111/j.1467-8624.1996.tb01791.x

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*, 57-75. http://dx.doi.org/10.1080/0729436990180105

Birenbaum, M. (1997). Assessment preferences and their relationship to learning strategies and orientations. *Higher Education, 33*, 71-84. https://doi.org/10.1023/A:1002985613176

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918-928. http://dx.doi.org/10.1037/0278-7393.34.4.918

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281. https://doi.org/10.3102/00346543065003245

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14. doi:10.1111/j.2044-8279.1988.tb00874.x

Cassidy, T. (2013). *Environmental psychology: Behaviour and experience in context*. Hove and New York: Psychology Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438-481. https://doi.org/10.3102/00346543058004438

Crosby, J. R., & Monin, B. (2007). Failure to warn: How student race affects warnings of potential academic difficulty. *Journal of Experimental Social Psychology, 43,* 663-670. https://doi.org/10.1016/j.jesp.2006.06.007

Dee, T. S., Dobbie, W., Jacob, B. A., & Rockoff, J. (2016). *The causes and consequences of test score manipulation: Evidence from the new york regents examinations.* Stanford, CA: Stanford Center for Education Policy Analysis. Retrieved from http://cepa.stanford.edu/wp16-08.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Thousand Oaks, CA: Sage Publications, Inc.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22,* 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4-58. https://doi.org/10.1177/1529100612453266

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Flavell, J. H. (1971). First discussant's comments: What is memory development the development of? *Human Development, 14*, 272-278. doi:10.1159/000271221

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231-235). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive judgments: a heuristic account. *Journal of Cognitive Psychology, 29,* 184-201. http://dx.doi.org/10.1080/20445911.2016.1229669

Goodman, J. S., Wood, R. E., & Chen, Z. (2011). Feedback specificity, information processing, and transfer of training. *Organizational Behavior and Human Decision Processes, 115*, 253-267. doi:10.1016/j.obhdp.2011.01.001

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160-170. doi:10.1037//0022-0663.92.1.160

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81-112. https://doi.org/10.3102/003465430298487

Henzi, D., Davis, E., Jasinevicius, R., & Hendricson, W. (2006). North American dental students' perspectives about their clinical education. *Journal of Dental Education, 70,* 361-377. Retrieved from http://www.jdentaled.org

Kelley, C., & Jacoby, L. (1996). Memory attributions: Remembering, knowing and feeling of knowing. In L.M. Reder (Eds.), *Implicit Memory and Metacognition* [Kindle version]. Retrieved from https://www.amazon.com.au/.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93,* 329-343. http://dx.doi.org/10.2307/1422236

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284. http://dx.doi.org/10.1037/0033-2909.119.2.254

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences, 25,* 67-72. doi: 10.1016/j.lindif.2013.01.005

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349-370. http://dx.doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9,* 149-171. https://doi.org/10.1006/ccog.2000.0433

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289-325). Cambridge:: Cambridge University Press.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36-39. http://dx.doi.org/10.1037/0096-3445.135.1.36

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*, 147-162. http://dx.doi.org/10.1037/0096-3445.131.2.147

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology: Applied, 19*, 1-13. doi:10.1037/a0032147

Kuh, G. D., & Hu, S. (1999). Unraveling the complexity of the increase in college grades from the mid-1980s to the mid-1990s. *Educational Evaluation and Policy Analysis, 21*, 297-320. https://doi.org/10.3102/01623737021003297

Lam, C. F., DeRue, D. S., Karam, E. P., & Hollenbeck, J. R. (2011). The impact of feedback frequency on learning and task performance: Challenging the "more is better" assumption. *Organizational Behavior and Human Decision Processes, 116*, 217-228. https://doi.org/10.1016/j.obhdp.2011.05.002

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183. doi:10.1016/0030-5073(77)90001-0

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537-563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1263-1274. http://dx.doi.org/10.1037/0278-7393.21.5.1263

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159-163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology, 68*, 465-489. https://doi.org/10.1146/annurev-psych-010416-044022

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*, 267-271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*, 325-335. doi:10.1080/09658219408258951

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of Metacognitive Judgments in the Allocation of Study During Multitrial Learning. *Psychological Science, 5*, 207-213. doi:10.1111/j.1467-9280.1994.tb00502.x

Nelson, T. O., & Narens. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125-173. doi:10.1016/S0079-7421(08)60053-5

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied, 17*, 288-302. doi:10.1037/a0024749

Razran, G. (1955). A note on second-order conditioning and secondary reinforcement. *Psychological Review, 62*, 327-332. doi:10.1037/h0047135

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record, 114*, 1-23. Retrieved from https://rampages.us/

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review, 3*, 16-20. https://doi.org/10.1007/BF02322211

Rosenthal, R., & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis (2nd ed.)*. New York: McGraw-Hill.

Sadler, R. D. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*, 175-194. http://dx.doi.org/10.1080/0260293042000264262

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124-128. doi:10.1037/0096-3445.134.1.124

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology, 6*, 147-151. doi:10.1027/1614-2241/a000016

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189. https://doi.org/10.3102/0034654307313795

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270-277. http://dx.doi.org/10.1037/h0062535

Staddon, J. (2014). *The new behaviorism*: [Kindle version]. Retrieved from Amazon.com

Steenkamp, N., & Roberts, R. (2017). Unethical practices in response to poor student quality: An Australian perspective. *The Accounting Educators' Journal, 26,* 89-119. Retrieved from http://www.aejournal.com

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66-73. doi:10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024-1037. http://dx.doi.org/10.1037/0278-7393.25.4.1024

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*.1124-1131. doi:10.1126/science.185.4157.1124

Vinton, L., & Wilke, D. J. (2011). Leniency bias in evaluating clinical social work student interns. *Clinical Social Work Journal, 39*, 288-295. https://doi.org/10.1007/s10615-009-0221-5

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet, 357*, 945-949. doi:10.1016/S0140-6736(00)04221-5

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on*

Survey Research Methods, 354-359. American Statistical Association, Retrieved from https://eric.ed.gov

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611-617. http://dx.doi.org/10.1037/0033-2909.110.3.611

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29*, 663-676. https://doi.org/10.3102/00028312029003663

Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education, 25*, 1081-1094. doi:10.1080/0950069032000052207

Appendices

Appendix A: Ethics Approval

Ethics Amendment Approved: H0012660 Confidence in memory

Katherine Shaw

Wed 15/03/2017 3:33 PM

To: Matt Palmer <matt.palmer@utas.edu.au>;

Cc: Jim Sauer <jim.sauer@utas.edu.au>; Andrew Heathcote <andrew.heathcote@utas.edu.au>; Nicole.mccallum@flinders.edu.au <Nicole.mccallum@flinders.edu.au>; Frances Parkes <fam.parkes@utas.edu.au>; Matthew Gretton <matthew.gretton@utas.edu.au>; Rachel Breen <rjbreen@utas.edu.au>; Laura Brumby <lebrumby@utas.edu.au>; Roderick Garton <roderick.garton@utas.edu.au>; Caitlin Gleeson <clg0@utas.edu.au>; Valera Griffin <valerag@utas.edu.au>; Glenys Holt <glenys.holt@utas.edu.au>; Amelia Kohl <akohl@utas.edu.au>; Rafal Kozlowski <rafalk@utas.edu.au>; Talira Kucina <talira.kucina@utas.edu.au>; Morgan Norris <mnorris0@utas.edu.au>; Terry Purton <terry.purton@utas.edu.au>; Miriam Rainsford <miriam.rainsford@utas.edu.au>; Daniel Zuj <daniel.zuj@utas.edu.au>;

Dear Dr Palmer

Ethics Ref No: H0012660
Project title: Confidence in memory

This email is to confirm that the following amendment was approved by the Chair of the Tasmania Social Sciences Human Research Ethics Committee on 15/3/2017:

- Addition of student researchers Mr Rod Garton, Ms Amelia Kohl, Ms Morgan Norris, Mr Rafal Kozlowski, Ms Talira Kucina and Ms Rachel Breen.
- Removal of student researchers Ms Rebecca Healy, Ms Kate Edwards, Ms Catherine Bishop, Ms Katie-Lee Crawford, Ms Katie Henderson, Ms Rebecca Kaiser, Mr Robert Kirkis, Mr Michael O'Leary and Mr Tane Thomas.

All committees operating under the Human Research Ethics Committee (Tasmania) Network are registered and required to comply with the National Statement on Ethical Conduct in Human Research (NHMRC 2007, updated May 2015).

This email constitutes official approval. If your circumstances require a formal letter of amendment approval, please let us know.

Should you have any queries please do not hesitate to contact me.

Kind regards
Katherine

Katherine Shaw
Executive Officer, Social Sciences HREC
Office of Research Services | Research Division
University of Tasmania
Private Bag 1
Hobart TAS 7001
T +61 3 6226 2763
[www.utas.edu.au/research]www.utas.edu.au/research

UNIVERSITY of TASMANIA
CRICOS 00586B

Appendix B: Information Sheet

Locked Bag 30 Hobart
Tasmania 7001 Australia
Phone 0448439378
akohl@utas.edu.au

UTAS

The Accuracy of Judgments of Learning in Studying Swahili

Information Sheet for Participants

1. Invitation

I would like to invite you to participate in a psychology experiment about the accuracy of memory judgments. The experiment is being conducted by Rafal Kozlowski, supervised by Jim Sauer of the School of Psychology at the University of Tasmania.

2. What is the purpose of this study?

This experiment investigates the accuracy of memory judgments, and individuals' ability to assess their learning progress.

3. Why have I been invited to participate?

You may have been invited for a number of reasons. You may have been approached through the first year student participation pool. You may have previously added your name to a database of people interested in participating in our research. You may have responded to an advertisement. We are interested in testing a diverse range of people, so there is no specific reason why you have been invited.

Participation in this study is voluntary – you are entirely free to choose to participate or not, and there will be no consequences if you decide not to participate. If you do participate, any information you provide will be anonymous and no participants in the experiment will be individually identifiable.

4. What will I be asked to do?

Participation will take approximately 45 minutes (though may take a little longer). You will be presented with a variety of facial images to study, and will then be asked to make recognition judgements about these stimuli. The experiment is conducted entirely on computer.

5. Are there any possible benefits from participation in this study?

Participation in this research project will expose you to psychological research and will expand your understanding of how research is conducted in a laboratory setting. Furthermore, this research could contribute to the body of knowledge in applied memory (e.g., eyewitness memory), and our understanding of people's ability to assess the reliability of their own memory judgements.

6. Are there any possible risks from participation in this study?

There are no known physical, psychological, economic or social risks associated with your participation in this study. You will be assigned a randomly generated Participant ID Number, and thus your data will not be able to be linked to your identity.

7. What if I change my mind during or after the study?

That's fine - you are free to withdraw from the study at any time, and without providing an explanation. If you choose to withdraw during the study, your responses will be destroyed.

Locked Bag 30 Hobart
Tasmania 7001 Australia
Phone 0448439378
akohl@utas.edu.au

UTAS

UNIVERSITY OF TASMANIA

8. What will happen to the information when this study is over?

The data from this study will be kept in secure storage on the University of Tasmania premises for a period of five years after any publications (e.g., in academic journals) that involve the data. After this period, the data will be archived. Only the researchers will have access to the raw data.

The data will be stored anonymously. All responses will be anonymous and no identifying information will be collected from participants.

9. How will the results of the study be published?

The results of the study will be published in Rafal Kozlowski's honors thesis.

No individual participants will be identifiable in the publication of the results.

10. What if I have questions about this study?

If you have any questions about this study, please feel free to contact Rafal Kozlowski via phone on 0407 495 222 or email: akohl@utas.edu.au

This study has been approved by the Tasmanian Social Sciences Human Research Ethics Committee. If you have concerns or complaints about the conduct of this study, please contact the Executive Officer of the HREC (Tasmania) Network on (03) 6226 7479 or email human.ethics@utas.edu.au. The Executive Officer is the person nominated to receive complaints from research participants. Please quote ethics reference number H12660.

This information sheet is for you to keep. If you would like to participate in this study, please ask the researcher for a Consent Form to complete.

Thank you for your attention - your time is very much appreciated.

Appendix C: Consent Form

Locked Bag 30 Hobart
Tasmania 7001 Australia
Phone 0448439378
akohi@utas.edu.au

UTAS

UNIVERSITY OF TASMANIA

The Accuracy of Judgments of Learning in Studying Swahili

Participant Consent Form

1. I agree to take part in the research study named above.
2. I have read and understood the Information Sheet for this study.
3. The nature and possible effects of the study have been explained to me.
4. I understand that the study involves viewing a series of stimuli and answering questions about them.
5. I understand that participation involves no foreseeable risks.
6. I understand that all research data will be securely stored on the University of Tasmania premises for five years from the publication of the study results, and will then be destroyed unless I give permission for my data to be archived.

   I agree to have my study data archived. (Note that your data will be stored anonymously.)

   Yes ☐ No ☐
7. Any questions that I have asked have been answered to my satisfaction.
8. I understand that the researchers will maintain confidentiality and that any information I supply to the researcher will be used only for the purposes of the research.
9. I understand that the results of the study will be published so that I cannot be identified as a participant.
10. I understand that my participation is voluntary and that I may withdraw at any time without any effect.

    I understand that I will not be able to withdraw my data after completing the experiment as my data will be anonymous.

Participant's name: _____

Participant's signature: _____

Date: _____

Locked Bag 30 Hobart
Tasmania 7001 Australia
Phone 0448439378
akohl@utas.edu.au

UTAS

UNIVERSITY OF TASMANIA

Statement by Investigator

☐ I have explained the project and the implications of participation in it to this volunteer and I believe that the consent is informed and that he/she understands the implications of participation.

If the Investigator has not had an opportunity to talk to participants prior to them participating, the following must be ticked.

☐ The participant has received the Information Sheet where my details have been provided so participants have had the opportunity to contact me prior to consenting to participate in this project.

Investigator's name: _____

Investigator's signature: _____

Date: _____

Appendix D: Initial Debrief Sheet

Participant no:_____

Initial Debrief

Study:	**The Accuracy of Judgments of Learning in Studying Swahili**
Researcher:	Rafal Kozlowski, Psychology Honours Student, University of Tasmania, rafalk@utas.edu.au

What were the aims of this study?
This study is investigating the accuracy of memory judgments for studied items. In order to preserve the scientific rigour of the research (by ensuring that future participants remain naive to the purpose of the experiment and the full experimental hypotheses) we will not be providing a full debrief at this time. However, we will provide a full debrief as soon as data collection is complete.

You can also obtain a summary of the results of the study by writing an email to Rafal Kozlowski using the contact information above. We expect that such a summary will be available by late October 2017. It will be emailed to you automatically if you enter your email address into our results request list.

If you have any questions about the study please ask the experimenters, they will be happy to answer them now (although they are unable to reveal the exact purpose/hypotheses of the experiment).

If, for any reason, you wish to withdraw your data once you have left you can do this by writing an email to this effect to Rafal Kozlowski using the contact information provided above and quoting your participant number at the top of this sheet.